Exploring Compositional Generalization of Multimodal LLMs for Medical Imaging

Zhenyang Cai[†], **Junying Chen**[†], **Rongsheng Wang**[†], **Weihong Wang**, **Yonglin Deng**, **Dingjie Song**, **Yize Chen**, **Zixu Zhang**, **Benyou Wang**[*]
The Chinese University of Hong Kong, Shenzhen
wangbenyou@cuhk.edu.cn

Abstract

Medical imaging provides essential visual insights for diagnosis, and multimodal large language models (MLLMs) are increasingly utilized for its analysis due to their strong generalization capabilities; however, the underlying factors driving this generalization remain unclear. Current research suggests that multi-task training outperforms single-task as different tasks can benefit each other, but they often overlook the internal relationships within these tasks. To analyze this phenomenon, we attempted to employ *compositional generalization* (CG), which refers to the models' ability to understand novel combinations by recombining learned elements, as a guiding framework. Since medical images can be precisely defined by **M**odality, **A**natomical area, and **T**ask, naturally providing an environment for exploring CG, we assembled 106 medical datasets to create **Med-MAT** for comprehensive experiments. The experiments confirmed that MLLMs can use CG to understand unseen medical images and identified CG as one of the main drivers of the generalization observed in multi-task training. Additionally, further studies demonstrated that CG effectively supports datasets with limited data and confirmed that MLLMs can achieve CG across classification and detection tasks, underscoring its broader generalization potential. Med-MAT is available at https://github.com/FreedomIntelligence/Med-MAT.

1 Introduction

Medical imaging provides essential visual insights into the structures of the human body, making it a critical tool for medical diagnosis. Recently, multimodal large language models (MLLMs) (Liu et al., 2023; Li et al., 2024; Chen et al., 2024b) have been employed to analyze these images due to their strong interpretability and generalization capabilities. In this paper, we focus on the latter: generalization of MLLMs in medical imaging. Current research (Mo and Liang, 2024; Ren et al., 2024) has demonstrated that models trained on multiple tasks outperform those trained on a single task as they can leverage potential knowledge from other tasks. Yet, the underlying factors that contribute to this generalization remain insufficiently explored.

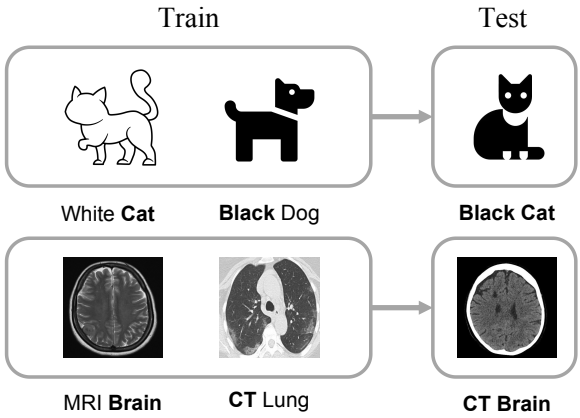

Figure 1: Examples of *Compositional Generalization*: The model is required to understand unseen images by recombining the fundamental elements it has learned.

To this end, we take the perspective of *composition generalization* (CG) (Li et al., 2019; Xu et al., 2022; Tang et al., 2024) to investigate the generalization phenomenon of mutual improvement in MLLMs' understanding of medical images. Specifically, CG is the model's ability to learn fundamental elements and recombine them in novel ways to understand unseen combinations (e.g., learning ***Cat*** from *White* ***Cat*** and ***Black*** from ***Black*** *Dog*, then generalizing to ***Black Cat***, as shown in Figure 1).

In this paper, we categorize each image to three elements: **M**odality, **A**natomical area, and medical **T**ask, presenting numerous natural opportunities for CG. We defined these three elements as the **MAT-Triplet** and collected 106 medical datasets, subsequently merging those that share the same *MAT-Triplet* to create the **Med-MAT** dataset.

[†]Equal Contribution. [*]Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13057–13079
July 27 - August 1, 2025

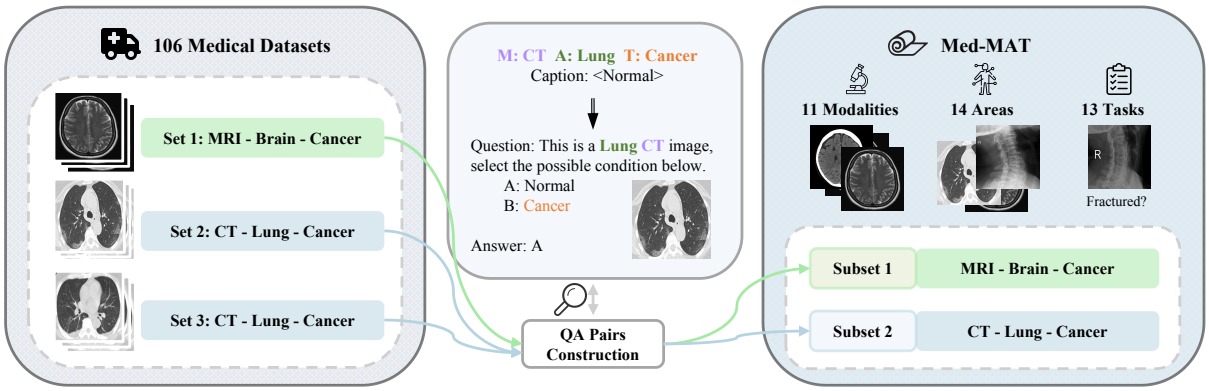

Figure 2: The process of integrating a vast amount of labeled medical image data to create Med-MAT.

Ultimately, Med-MAT comprises 53 subsets, encompassing 11 modalities, 14 anatomical regions, and 13 medical tasks, providing a foundation for investigating CG and other generalization methods.

To verify the existence of CG, we designated specific datasets as ***Target*** data and selected all ***Related*** data from Med-MAT that shared the same MAT-Triplet with the *Target* data. Using these data combinations, we accessed the generalization performance of MLLMs and observed that they could leverage CG to understand unseen medical images. To further validate this finding, we repeated the experiments on different MLLMs and obtained consistent results, confirming the universality of CG.

Building on these insights, we expanded the number of combinations and observed the changes in model generalization performance after deliberately disrupting CG, ultimately revealing that CG is a key factor driving the generalization of MLLMs. Furthermore, we explored the potential applications of CG and its performance across classification and detection tasks, finding that CG enhances MLLMs' ability to handle medical scenarios with limited training data and improves their capacity for spatial awareness.

Here are the key contributions of our work: 1) A VQA dataset, Med-MAT, has been constructed, providing a platform to explore the generalization of MLLMs on medical images. 2) Through this dataset, we observed that MLLMs in different architectures can utilize compositional generalization to understand unseen images and demonstrated that this is one of the main forms of generalization for medical MLLMs. 3) Finally, the real-world applicability of CG, along with its presence across detection and classification tasks, has been further explored, highlighting its potential to enhance data-efficient training and its broad applicability.

2 A Pilot Study on Generalization

2.1 Data Collection (Med-MAT)

Most existing datasets for MLLMs (Zhang et al., 2023c; Li et al., 2024; Chen et al., 2024b), primarily VQA datasets, provide broad coverage but lack attribute annotations for individual samples, which are not suitable for CG exploration. To address this gap, we curated a large collection of image-text pairs to develop **Med-MAT**, ensuring that each sample is explicitly defined by MAT-Triplet.

Data Construction Med-MAT contains a total of 106 image-label pair medical datasets, sourced from various medical public challenges or high-quality annotated datasets. All datasets are categorized according to their MAT-Triplet, with data having identical elements grouped into a single subset (Figure 2). Labels are manually clustered to ensure that annotations with the same meaning are not repeatedly used. In total, Med-MAT covers 11 medical modalities, 14 anatomical areas, and 13 medical tasks, hoping that it can spread across various medical tasks like a mat. (Data lists are shown in Appendix B)

Data Distribution All subsets are divided into training and test sets following their original distributions or using a 9:1 ratio. To ensure a fair comparison, each training set is limited to 3,000 samples [1], with label balance maintained as much as possible. Any subset that cannot meet this requirement is treated as an OOD (out-of-distribution) dataset. For the test sets, we strictly balance the number of samples per label to ensure that the accuracy metric reliably reflects model performance.

QA Pairs Construction To enable MLLMs to directly train and test on Med-MAT, all image-label

[1] Most datasets contain around 3,000 samples.

Subset No.	02	03	07	08	09	11	13	14	15	16	18	19	21	22	23	25	26	28	30	31	32	33	35	36	37
Baseline	21	47	40	25	26	27	28	24	22	24	25	23	49	26	25	24	49	30	49	21	49	20	25	23	19
Single-task Training	24	49	50	68	65	76	83	53	61	32	29	26	57	53	28	24	57	64	89	60	97	54	29	51	49
Multi-task Training	**96**	**89**	**80**	**80**	**79**	**97**	**92**	**88**	**76**	**57**	**88**	**74**	**87**	**86**	**93**	**52**	**98**	**72**	**94**	**61**	**100**	**72**	**75**	**60**	**50**

Table 1: Accuracy(%) of different models on In-Distribution datasets (each dataset contains over 3,000 samples, with 3,000 selected for training). Within each segment, **bold** highlights the best scores, and underline indicates the second-best. *Baseline* represents the results without any training, *Single-task Training* refers to the results after training on a single dataset, and *Multi-task Training* represents the results after training on all datasets.

Subset No.	01	04	05	06	10	12	17	20	24	27	29	34
Baseline	32	25	33	**33**	48	27	33	13	34	37	31	20
Multi-task Training	**39**	**26**	**70**	31	**58**	**38**	**61**	**40**	**35**	**41**	**55**	**50**

Table 2: Accuracy(%) of different models on Out-Of-Distribution Dataset (each dataset contains fewer than 3,000 samples and is used only for testing). **Bold** highlights the best scores. *Multi-task Training* represents the results after training on all datasets.

paired data were converted into a visual question-answering (VQA) format (Figure 3). Specifically, each subset was manually assigned 6 instructions to guide the MLLM in answering the subset task. For convenience, all samples were converted into single-choice questions with up to four options, and the remaining distractor options were randomly drawn from other labels within the subset. To mitigate potential evaluation biases arising from varying option counts, the ImageWikiQA dataset (Zhang et al., 2024b), a non-medical dataset consisting of single-answer, four-option questions, was incorporated during the training.

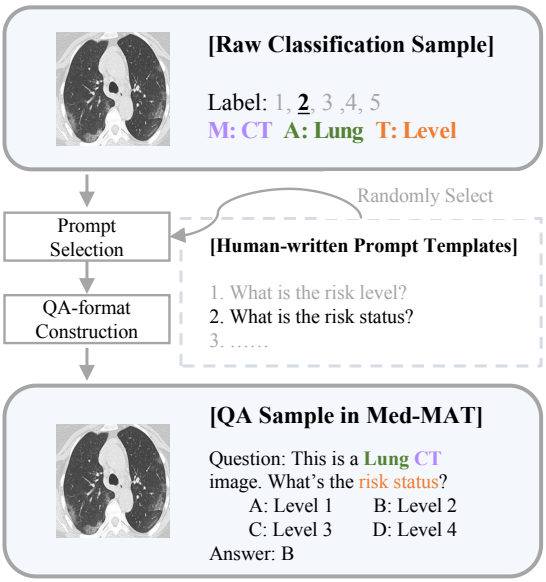

Figure 3: An example of formatting a raw classification sample into a Question-answering sample in Med-MAT.

2.2 Observation

Experiment Setup We chose LLaVA-v1.5-7B-Vicuna (Liu et al., 2023) as the base model due to its transparent pretraining process and minimal use of medical data, reducing the risk of knowledge leakage. Leveraging MLLM's flexibility, we enabled task switching and generalization by adjusting prompts, streamlining generalization studies. Each experiment ran for 5 epochs on 8 A800 GPUs with a batch size of 32 and a learning rate of 5e-6.

Analysis To access the generalization of MLLMs, we trained the baseline on all ID datasets to simulate *Multi-task Training* and separately trained on individual ID datasets to establish the *Single-task Training* as the control group. We then evaluated the models on all datasets. The results in Table 1 and 2 confirm that *Multi-task Training* outperformed *Single-task Training* on specific tasks and improved OOD prediction, suggesting certain data combinations enhance classification and identifying valuable combinations for medical tasks warrants further research. This observation leads to a research question (***RQ***):

> *What drives the generalization observed in MLLMs during Multi-task Training?*

To address it, we aim to explore the generalization mechanism of MLLMs from the perspective of compositional generalization (CG).

3 Proof of Concept on CG

This section will prove the existence of CG in MLLMs, offering preliminary insights to address the ***RQ*** and providing support for further analysis.

3.1 Experiment Setup

To explore the existence of CG from a finer perspective, this section focuses on CG with only two

Related Combination				Target Subset		Baseline	Baseline+	Trained	CG Helps
Lung	COVID	Brain	**Cancer**	**Lung**	**Cancer**	25	25	27	✓
Lung	Cancer	Brain	**State**	**Lung**	**State**	47	46	50	✓
Brain	Cancer	Lung	**State**	**Brain**	**State**	33	50	57	✓
Bones	Level	Lung	**State**	**Bones**	**State**	49	53	51	✗
Bones	Level	Brain	**State**	**Bones**	**State**	49	53	72	✓
Bones	Level	Breast	**Diseases**	**Bones**	**Diseases**	37	33	39	✓
Bones	Level	Lung	**Diseases**	**Bones**	**Diseases**	37	33	43	✓
Bones	Level	Chest	**Diseases**	**Bones**	**Diseases**	37	31	43	✓
Bones	State	Breast	**Diseases**	**Bones**	**Diseases**	37	37	43	✓
Bones	State	Lung	**Diseases**	**Bones**	**Diseases**	37	37	43	✓
Bones	State	Chest	**Diseases**	**Bones**	**Diseases**	37	37	41	✓
Lung	COVID	Breast	**Diseases**	**Lung**	**Diseases**	49	48	51	✓
Lung	COVID	Bones	**Diseases**	**Lung**	**Diseases**	49	48	52	✓
Lung	COVID	Chest	**Diseases**	**Lung**	**Diseases**	49	48	51	✓
CT	Cancer	X-ray	**COVID**	**CT**	**COVID**	47	46	72	✓
X-ray	Diseases	CT	**COVID**	**X-ray**	**COVID**	30	21	49	✓
X-ray	Diseases	CT	**State**	**X-ray**	**State**	30	21	46	✓
CT	State	X-ray	**Cancer**	**CT**	**Cancer**	33	28	28	✗
X-ray	Bones	CT	**Brain**	**X-ray**	**Brain**	49	49	91	✓
X-ray	Lung	CT	**Brain**	**X-ray**	**Brain**	49	50	81	✓
X-ray	Bones	CT	**Brain**	**X-ray**	**Brain**	25	51	74	✓
X-ray	Lung	CT	**Brain**	**X-ray**	**Brain**	49	52	52	✗
CT	Lung	X-ray	**Brain**	**CT**	**Brain**	33	50	60	✓
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	25	25	36	✓
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	47	50	81	✓
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	47	50	71	✓
X-ray	Bones	CT	**Lung**	**X-ray**	**Lung**	30	32	28	✗
X-ray	Brain	CT	**Lung**	**X-ray**	**Lung**	30	32	35	✓
X-ray	Bones	CT	**Lung**	**X-ray**	**Lung**	30	32	41	✓
X-ray	Brain	CT	**Lung**	**X-ray**	**Lung**	30	32	42	✓
Der - Skin	Cancer	FP - Fundus	**Diseases**	**Der - Skin**	**Diseases**	25	29	33	✓
Der - Skin	Cancer	OCT - Retine	**Diseases**	**Der - Skin**	**Diseases**	25	29	33	✓
Der - Skin	Diseases	DP - Mouth	**Cancer**	**Der - Skin**	**Cancer**	40	33	63	✓
Der - Skin	Diseases	Mic - Cell	**Cancer**	**Der - Skin**	**Cancer**	40	33	63	✓
DP - Mouth	State	Der - Skin	**Cancer**	**DP - Mouth**	**Cancer**	48	50	52	✓
DP - Mouth	State	Mic - Cell	**Cancer**	**DP - Mouth**	**Cancer**	48	50	55	✓
FP - Fundus	Diseases	Mic - Cell	**Level**	**FP - Fundus**	**Level**	33	36	42	✓
Mic - Cell	Recognition	FP - Fundus	**Level**	**Mic - Cell**	**Level**	23	33	32	✗
Mic - Cell	Recognition	Der - Skin	**Cancer**	**Mic - Cell**	**Cancer**	49	50	50	✗
Mic - Cell	Recognition	DP - Mouth	**Cancer**	**Mic - Cell**	**Cancer**	49	51	62	✓
Mic - Cell	Level	Der - Skin	**Cancer**	**Mic - Cell**	**Cancer**	49	51	52	✓
Mic - Cell	Level	DP - Mouth	**Cancer**	**Mic - Cell**	**Cancer**	49	51	58	✓
Mic - Cell	Cancer	FP - Fundus	**Level**	**Mic - Cell**	**Level**	23	24	27	✓

Table 3: Generalization results on classification datasets: *Related Combination* is the training set, *Target Subset* is the goal. *Baseline*, *Baseline+*, and *Trained* represent the model's accuracy(%) without training, trained on randomly sampled unrelated data, and trained on related data, respectively. ✓ in *CG Helps* indicates successful generalization, while ✗ denotes failure. The 4 segmented areas represent different Direction Types: fixed modality, fixed area, fixed task, and modality-area paired combinations. Although some combinations share the same name, they differ because they fix different elements.

MAT-Triplet elements varying while the third remains constant. Additionally, we identified specific Modality-Area pairs, such as dermoscopy paired consistently with skin, which were treated as a special category. These 4 different fixed formats were classified into distinct ***Direction Types***.

We adhered to the training setup described in Section 2.2 and evaluated the model's performance on the *Target* data. *Baseline* refers to the model without any training, while *Trained* refers to the model trained solely on *Related* data. To ensure that our conclusions are not influenced by the amount of training data, we randomly sampled an equal number of data from the *Unrelated* subsets, and this configuration is referred to as *Baseline+*.

3.2 Results

Results are shown in Table 3 and it can be observed that almost all CG combinations are able to generalize to downstream tasks, highlighting that MLLMs can leverage CG to generalize *Target* data across all Direction Types. Besides that, since this experiment focused solely on two-element tuples, we further investigated three-element tuples in Appendix A.4, where we also observed similarly strong generalizations when obtaining MAT-Triplet elements from three different datasets.

Take-away 1: *MLLMs can leverage CG to understand unseen medical images.*

In the *Baseline+* setting, we removed all datasets sharing any MAT-Triplet element with the *Target*

Related Combination				Target Subset		Qwen	Llama
Bones	State	Breast	**Diseases**	**Bones**	**Diseases**	+4	+7
Lung	COVID	Bones	**Diseases**	**Lung**	**Diseases**	+11	+11
X-ray	Diseases	CT	**COVID**	**X-ray**	**COVID**	+5	+5
X-ray	Diseases	CT	**State**	**X-ray**	**State**	+8	+8
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	+1	−2
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	+7	+8
FP - Fundus	Diseases	Mic - Cell	**Level**	**FP - Fundus**	**Level**	−3	+6
Mic - Cell	Recognition	FP - Fundus	**Level**	**Mic - Cell**	**Level**	+7	+22

Table 4: Result of Qwen2-VL and Llama-3.2-Vision on selected classification datasets in Med-MAT. *Qwen* and *Llama* represent the accuracy(%) gains they achieved on the respective backbones through CG.

data. Consequently, *Baseline+* models perform at near-random levels on the test set, indicating they failed to acquire target-relevant knowledge. This suggests that only datasets related through the MAT-Triplet can help the model learn and generalize to new target tasks.

Take-away 2: *Generalization arises in medical datasets in which at least partial MAT elements pre-exist during training.*

3.3 Extending CG to other Backbones

LLaVA was selected as the baseline because its training data and processes are publicly available, ensuring minimal exposure to medical images and preventing bias in the integration of medical image knowledge into the MLLM. To ensure that the results are not affected by the training data or the visual encoder of LLaVA, we randomly sampled two combinations from each *Direction Type* to investigate CG on Qwen2-VL-7B (Wang et al., 2024a) and Llama3.2-11B-Vision (Meta AI, 2024).

Qwen2-VL undergoes additional training on proprietary data based on ViT and incorporates a strategy to adjust the number of vision tokens according to resolution. Llama3.2-Vision, on the other hand, pretrains its own vision encoder from scratch using proprietary data. Thus, both models serve as a means to assess whether MLLMs with different training data and vision encoders can still leverage CG to understand unseen images, ensuring that CG is not merely an artifact of LLaVA's data fitting or specific to its vision encoder.

Table 4 presents the experimental results, showing that both selected backbones exhibit a certain degree of generalization across most tasks. This suggests that despite differences in pre-train data and vision encoders, different MLLMs can still leverage CG to understand unseen images.

Take-away 3: *CG persists across different MLLM backbones.*

4 Scaling Combination in CG

After confirming that CG is indeed a form of generalization in MLLMs, we expanded the number of participating combinations to explore the generalizability of CG and examine its relationship with the generalization exhibited by *Multi-task Training* to address the ***RQ***.

4.1 Experiment Setup

Two sub-questions have been defined to verify the applicability of CG in multiple data combinations and examine its role in *Multi-task Training*.

- **(Q1)** While previous experiment on CG indicated that *Unrelated* combinations provide no benefit to *Target* data, can generalization arise when training incorporates more *Unrelated* combinations, simulating a multi-task scenario?
- **(Q2)** Previous studies suggest that *Multi-task Training* generally promotes better generalization than single-task training. If the CG conditions in *Multi-task Training* are deliberately disrupted, will the resulting generalization effect be affected?

Selection Strategy To ensure a balanced evaluation of *Related* and *Unrelated* combinations, Subset 03 and Subset 28 were chosen as *Target* datasets because they exhibit the most balanced ratios of *Related* to *Unrelated* subsets (13:11 for Subset 03 and 11:13 for Subset 28), making them ideal for providing a diverse range of compositions in the scale-up experiments.

The baseline was trained on all subsets excluding the *Target* data to evaluate the claim that mixing multi-task data enhances generalization (*All Data*). To construct multiple comparative experiments, models were further trained on either *Related* or *Unrelated* subsets (*All Related* / *All Unrelated*) to address Q1. For

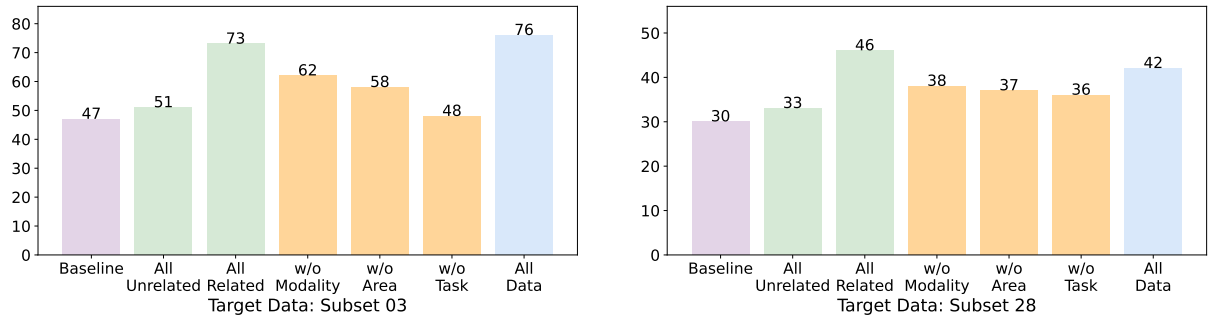

Figure 4: Accuracy(%) results on the *Target* dataset for various models. *All Related/Unrelated* models are trained on all the related or unrelated datasets of the *Target* Data. *w/o Modality/Area/Task* are trained on All Related datasets but omit those sharing the same element as the *Target* Data, to intentionally disrupt CG. *All Data* uses all available training sets. (Note: The *Target* Data is excluded from training to observe generalization.)

Q2, individual MAT-Triplet elements were systematically removed from the *Related* subsets (*Related w/o Modality / Area / Task*), disrupting CG and assessing the ability to maintain generalization. To ensure consistency, the total data volume in all experiments was limited to 15,000 samples, aligning with the number of ID subsets available after excluding related tasks from Subset 03.

4.2 Analysis of Scaling Experiment

Figure 4 illustrates the results. It can be observed that even when we expanded the *Unrelated* combination volumes and increased task diversity, the performance of *All Unrelated* remains close to the *Baseline*, indicating that these datasets can not support MLLMs to understand the *Target* data.

Take-away 4: *Datasets without MAT-Triplet overlap offer limited benefit for generalization even in the multi-task training scenario (Q1).*

Besides, *w/o Modality / Area / Task* showed significant accuracy drops compared to *All Related*, despite holding the training data volume constant. This indicates that if the CG combinations are forcibly disrupted, MLLMs will lose a significant amount of generalization capability for the target data.

Take-away 5: *Disrupting CG leads to a significant decline in generalization ability. (Q2).*

Notably, *All Related* achieves a performance level comparable to *All Data*, where all datasets are included in training. This suggests that CG plays a crucial role in enhancing the generalization effect of *Multi-task Training*. Therefore, in conclusion:

Take-away 6: *CG plays an important role in generalization for MLLMs in medical imaging.*

5 Potential Applications of CG

As MLLMs can use CG to generalize unseen medical images, this section attempts to explore its potential applications in training medical MLLMs.

5.1 Generalization without *Target* Data

In medical tasks, new and unpredictable conditions, like COVID-19, can emerge at any time. Exploring how to use CG to help MLLMs enhance their ability to identify unknown diseases in the absence of specific datasets is both important and meaningful.

We selected some *Target* datasets and trained the MLLMs using *Related* and *Unrelated* data to observe their generalization to the *Target* data. The generalization trend was assessed by progressively increasing the size of the combination datasets.

Selection Strategy To highlight the generalization trends, the combinations with strong generalization results were selected from the main experiments. For fairness, we chose the combinations across four types where *Trained* results exceed both *Baseline* and *Baseline+* by at least 10. If multiple combinations meet the criteria, a random seed of 42 was used to determine the selection.

Analysis The experimental results are shown in Figure 5, where the red line represents the accuracy curve for *Related* combinations, and the purple line shows the gain from *Unrelated* combinations. The *Related* combinations group significantly outperformed the *Unrelated* combinations in terms of generalization across all tasks, with this ability continuing to improve as the data size increased. This suggests that *Related* combinations, leveraging CG, enhance the model's ability to understand unknown medical tasks.

Take-away 7: *CG might enable MLLMs to handle tasks without dedicated training data.*

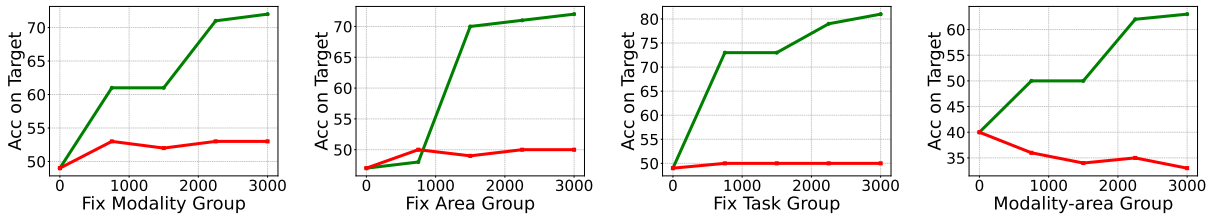

Figure 5: The accuracy curve reflects the impact of gradually increasing the composition dataset size without using *Target* data in training. The green and red lines represent training with Related and Unrelated Data, respectively.

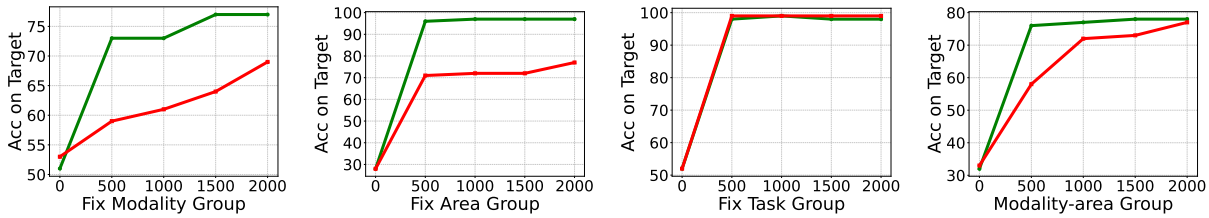

Figure 6: The accuracy curve shows the impact of increasing the composition dataset volume while incorporating *Target* data in training. The green and red lines represent training with Related and Unrelated Data, respectively.

5.2 Generalization with Limited *Target* Data

This section investigates the benefit of CG for tasks with limited data, e.g. processing medical images in rare conditions.

Selection Strategy To assess generalization in limited data scenarios, we select combinations with poor generalization from Table 3. Specifically, for each *Direction Type*, we randomly choose a CG combination with weak generalization (i.e., rows marked with ✗ in the last column of Table 3). For these combinations, we introduce an additional 2,000 examples from the *Target* data.

Analysis Figure 6 shows the results. It can be seen that as we gradually expand the training volume of *Target* data, adding the *Related* combination for training enabled the model to reach the peak performance more quickly. This suggests that leveraging CG to assist low-data medical scenarios can lead to more data-efficient training, even when CG does not directly result in significant generalization gains in these scenarios.

Take-away 8: *Although CG might not provide direct generalization gains, it helps data efficiency for MLLM training.*

6 CG across Detection and Classification

Previous studies (Ren et al., 2024; Wang et al., 2025) have shown that jointly training classification and detection tasks can mutually enhance their performance. Building on this, we investigate whether MLLMs can leverage classification data (e.g., visual knowledge) and detection data (e.g., spatial information) through CG to improve downstream classification (***Q1***) or detection tasks (***Q2***).

6.1 Experiment Settings

Training Setup Each generalization combination used for training in this experiment includes one detection dataset and one classification dataset to examine the generalization relationship between these two vision tasks. The detailed training parameters can be found in Appendix A.6.

Model Selection Next-Chat (Zhang et al., 2023a) and MiniGPT-v2 (Chen et al., 2023a) are selected as baselines, representing the two main approaches MLLMs use for detection tasks. The former treats bounding boxes as embeddings and decodes them into coordinates using a visual decoder, while the latter processes coordinate points as special text tokens and generates bounding box coordinates directly as output text.

Data Processing Med-MAT includes both detection and segmentation datasets. If a segmentation dataset provides object localization using masks, we extract the outermost coordinates of the corresponding mask to construct a bounding box, facilitating generalization experiments for detection. Subsequently, to streamline the experiments, we structured the dataset following the official data formats of Next-Chat and MiniGPT-v2.

6.2 Benefits for Classification (Q1)

In this experiment, all possible CG combinations were selected and the CG-trained model will be

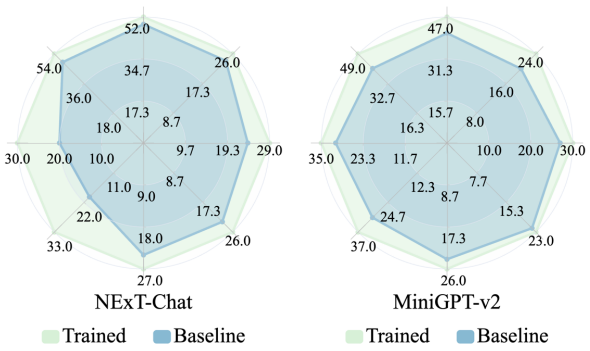

Figure 7: The accuracy(%) on Classification: Blue represents the untrained model, and green represents the CG-trained model. (details in Appendix A.5)

tested on classification task. The final results in Figure 7 show that all CG combinations demonstrated the model's successful utilization of detection data for CG to the *Target* data.

6.3 Benefits for Detection (Q2)

Subset 38 and 39 are selected as the objects in these datasets are relatively randomly distributed in the images, making them suitable for evaluating the model's detection capability. Subsequently, we selected certain classification datasets to construct CG for testing and used cIoU to evaluate the detection performance (follow (Chen et al., 2023a)).

Since both baselines lack localization capabilities for medical tasks, we incorporated a fixed amount of *Target* data into our experiments, adjusting the evaluation scenario to assess support in low-data settings. The results in Table 5 show that all selected CG combinations help MLLMs achieve better performance in detection tasks.

Related Combination		Target Subset	Next-Chat	MiniGPT-v2
D - Skin	C - Intestine	D - Intestine	+3.8	+4.1
D - Intestine	C - Skin	D - Skin	+8.4	+7.6

Table 5: *Next-Chat* and *MiniGPT-v2* respectively represent the cIoU gain brought by CG. C indicates classification task, D indicates detection task.

Take-away 9: *MLLMs can perform CG across classification and detection tasks.*

7 Related Work

Medical MLLMs Recently, adapting MLLMs to medical tasks has gained prominence due to their success in capturing complex visual features. Current MLLMs typically pair a visual encoder with a text-only LLM, aligning image data with language understanding. Such as Med-Flamingo (Moor et al., 2023) and Med-PaLM (Tu et al., 2024), fine-tuned general multimodal models and achieved notable results. Med-Flamingo enhanced OpenFlamingo-9B (Chen et al., 2024a) with medical data, while Med-PaLM adapted PaLM-E (Driess et al., 2023) using 1 million data points. Similarly, LLaVA-Med (Li et al., 2024), Med-Gemini (Saab et al., 2024), and HuatuoGPT-Vision (Chen et al., 2024b) utilized specialized datasets and instruction tuning to refine medical VQA tasks.

Generalization on Medical Imaging Generalization in medical imaging (Matta et al., 2024) has been extensively studied. Early methods utilized data manipulation techniques, such as data augmentation (Li et al., 2022; Zhang et al., 2022), to enhance model generalization on unseen medical data by adapting to varying distributions. Later approaches focused on representation learning (Le-Khac et al., 2020), preserving essential image information to enable models to handle more complex scenarios. Additionally, some studies (Ren et al., 2024) explore multiple aspects of medical image processing, examining how classification and segmentation tasks can mutually benefit each other.

Detection with MLLMs Recent studies employ various strategies to equip MLLMs with the capability to handle detection tasks, such as encoding regions as features to allow models to accept regions as input (Zhang et al., 2023b), representing object bounding box coordinates with text tokens (Wang et al., 2024c; Peng et al., 2023; Chen et al., 2023b), and employing unique identifiers for task instructions to improve learning efficiency. Additionally, some approaches introduce special tokens to represent images and use their hidden states to decode position information (Zhang et al., 2023a, 2024a).

8 Conclusion

To investigate whether MLLMs can leverage CG to generalize to unseen medical data, we constructed the Med-MAT dataset as a research platform for generalization experiments. The results confirmed the presence of CG and identified it as a key factor of MLLMs' generalization observed in multi-task learning. Further experiments showed that CG helps MLLMs handle limited data conditions, providing support for low-data medical tasks. Additionally, our findings showed that MLLMs can apply CG across detection and classification tasks, underscoring its broad generalization potential.

Limitations

The experiment confirms that MLLMs leverage CG for unseen medical images and data-efficient training. However, as shown in Section 4, disrupting CG reduces generalization but retains some effectiveness, indicating CG is just one aspect of MLLM generalization in medical imaging.

Potential Risks

Our research focuses on the compositional generalization of MLLMs on medical images, using data sourced from medical challenges and open-source datasets. However, further experiments are needed to mitigate potential risks when deploying this concept in real-world medical settings.

Acknowledgments

This work was supported by Shenzhen Medical Research Fund (No.C2406002) from the Shenzhen Medical Academy of Research and Translation (SMART), the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608), Shenzhen Science and Technology Program (Shenzhen Key Laboratory Grant No. ZDSYS20230626091302006), Shenzhen Stability Science Program 2023, and National Natural Science Foundation of China (NSFC) (72495131).

References

Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. 2020. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief*, 30:105474.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. 2020. Dataset of breast ultrasound images. *Data in brief*, 28:104863.

Shams Nafisa Ali, Md. Tazuddin Ahmed, Joydip Paul, Tasnim Jahan, S. M. Sakeef Sani, Nawshaba Noor, and Taufiq Hasan. 2022. Monkeypox skin lesion detection using deep learning models: A preliminary feasibility study. *arXiv preprint arXiv:2207.03342*.

Sharib Ali, Barbara Braden, Dominique Lamarque, Stefano Realdon, Adam Bailey, Renato Cannizzaro, Noha Ghatwary, Jens Rittscher, Christian Daul, and James East. 2020. Endoscopy disease detection and segmentation (edd2020).

MD Anouk Stein, Carol Wu, Chris Carr, George Shih, Jamie Dulkowski, kalpathy, Leon Chen, Luciano Prevedello, MD Marc Kohli, Mark McDonald, Peter, Phil Culliton, Safwan Halabi MD, and Tian Xia. 2018. Rsna pneumonia detection challenge. https://kaggle.com/competitions/rsna-pneumonia-detection-challenge. Kaggle.

Will Arevalo. 2020. Chexpert v1.0 small. https://www.kaggle.com/datasets/willarevalo/chexpert-v10-small. Kaggle.

A Asraf and Z Islam. 2021. Covid19, pneumonia and normal chest x-ray pa dataset. mendeley data v1 (2021).

Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. 2020. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis & Stereology*, 39(3):161–167.

Dev Batra. 2024. Fracture detection using x-ray images. https://www.kaggle.com/datasets/devbatrax/fracture-detection-using-x-ray-images. Kaggle.

Veronica Elisa Castillo Benítez, Ingrid Castro Matto, Julio César Mello Román, José Luis Vázquez Noguera, Miguel García-Torres, Jordan Ayala, Diego P Pinto-Roa, Pedro E Gardel-Sotomayor, Jacques Facon, and Sebastian Alberto Grillo. 2021. Dataset from fundus images for the study of diabetic retinopathy. *Data in brief*, 36:107068.

BenO, jljones, Kumar H, Meg Risdal, MRao, Vadim Sherman, Vipul, Wendy Kan, and Yau Ben-Or. 2017. Intel & mobileodt cervical cancer screening. https://kaggle.com/competitions/intel-mobileodt-cervical-cancer-screening. Kaggle.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. 2015. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111.

Bukun. 2019. Breast cancer histopathological database (breakhis). https://www.kaggle.com/datasets/ambarish/breakhis. Kaggle.

Olivia Cardozo, Verena Ojeda, Rodrigo Parra, Julio César Mello-Román, José Luis Vázquez Noguera, Miguel García-Torres, Federico Divina, Sebastian A. Grillo, Cynthia Villalba, Jacques Facon, Veronica Elisa Castillo Benítez, Ingrid Castro Matto, and Diego Aquino-Brítez. 2023. Dataset of fundus images for the diagnosis of ocular toxoplasmosis. *Data in Brief*, 48:109056.

Ling-Ping Cen, Jie Ji, Jian-Wei Lin, Si-Tong Ju, Hong-Jie Lin, Tai-Ping Li, Yun Wang, Jian-Feng Yang, Yu-Fen Liu, Shaoying Tan, et al. 2021. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature communications*, 12(1):4828.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2024a. Visual instruction tuning with polite flamingo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17745–17753.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, et al. 2024b. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Pingjun Chen. 2018. Knee osteoarthritis severity grading dataset. *Mendeley Data*, 1(10.17632).

Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. 2020. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE.

Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. 2019. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.

Will Cukierski. 2018. Histopathologic cancer detection. https://kaggle.com/competitions/histopathologic-cancer-detection. Kaggle.

Training Data. 2023. Computed tomography of the brain. https://www.kaggle.com/datasets/trainingdatapro/computed-tomography-ct-of-the-brain. Kaggle.

Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, Adrian Galdran, Miguel Ángel González Ballester, Gustavo Carneiro, Devika R G, Hrishikesh P S, Densen Puthussery, Hong Liu, Zekang Yang, Satoshi Kondo, Satoshi Kasai, Edward Wang, Ashritha Durvasula, Jónathan Heras, Miguel Ángel Zapata, Teresa Araújo, Guilherme Aresta, Hrvoje Bogunović, Mustafa Arikan, Yeong Chan Lee, Hyun Bin Cho, Yoon Ho Choi, Abdul Qayyum, Imran Razzak, Bram van Ginneken, Hans G. Lemij, and Clara I. Sánchez. 2023. Airogs: Artificial intelligence for robust glaucoma screening challenge. *arXiv preprint arXiv:2302.01738*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Fernando Feltrin. 2022. Brain tumor mri images 17 classes. https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-17-classes. Kaggle.

Mohammad Fraiwan, Ziad Audat, Luay Fraiwan, and Tarek Manasreh. 2022. Using deep transfer learning to detect scoliosis and spondylolisthesis from x-ray images. *Plos one*, 17(5):e0267851.

Huazhu Fu, Fei Li, José Ignacio Orlando, Hrvoje Bogunović, Xu Sun, Jingan Liao, Yanwu Xu, Shaochong Zhang, and Xiulan Zhang. 2019. Palm: Pathologic myopia challenge.

Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov. 2015. Med-node: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications*, 42(19):6578–6585.

Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. 2021. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 257–261. IEEE.

Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Fei Chen, and Guanbin Li. 2022. Thyroid region prior guided attention for ultrasound segmentation of

thyroid nodules. *Computers in Biology and Medicine*, 106389:1–12.

Shivanand Gornale and Pooja Patravali. 2020. Digital knee x-ray images.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828.

David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. 2016. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*.

Saba Hesaraki. 2022. Breast ultrasound images dataset (busi). https://www.kaggle.com/datasets/sabahesaraki/breast-ultrasound-images-dataset. Kaggle.

Md Nazmul Islam, Mehedi Hasan, Md Kabir Hossain, Md Golam Rabiul Alam, Md Zia Uddin, and Ahmet Soylu. 2022a. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. *Scientific Reports*, 12(1):1–14.

Towhidul Islam, Mohammad Arafat Hussain, Forhad Uddin Hasan Chowdhury, and B M Riazul Islam. 2022b. A web-scrapped skin image database of monkeypox, chickenpox, smallpox, cowpox, and measles. *bioRxiv 2022.08.01.502199*.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. 2014. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer.

Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. 2022. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific data*, 9(1):475.

JR2NGB. 2019. Cataract dataset. https://www.kaggle.com/datasets/jr2ngb/cataractdataset. Kaggle.

Nur Karaca. 2022. Nlm montgomery cxr set. https://www.kaggle.com/datasets/nurkaraca/nlm-montgomerycxrset. Kaggle.

Karthik, Maggie, and Sohier Dane. 2019. Aptos 2019 blindness detection. https://kaggle.com/competitions/aptos2019-blindness-detection. Kaggle.

Andrey Katanskiy. 2019. Skin cancer isic. https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic. Kaggle.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 2018. 100,000 histological images of human colorectal cancer and healthy tissue.

Daniel Kermany. 2018. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*.

Felipe Campos Kitamura. 2018. Head ct - hemorrhage.

Jorge F Lazo, Benoit Rosa, Michele Catellani, Matteo Fontana, Francesco A Mistretta, Gennaro Musi, Ottavio de Cobelli, Michel de Mathelin, and Elena De Momi. 2023. Semi-supervised bladder tissue classification in multi-domain endoscopic images. *IEEE Transactions on Biomedical Engineering*.

Trang Le, Casper F Winsnes, Ulrika Axelsson, Hao Xu, Jayasankar Mohanakrishnan Kaimal, Diana Mahdessian, Shubin Dai, Ilya S Makarov, Vladislav Ostankovich, Yang Xu, et al. 2022. Analysis of the human protein atlas weakly supervised single-cell classification competition. *Nature methods*, 19(10):1221–1229.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934.

Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9.

Sangjune L Lee, Poonam Yadav, Yin Li, Jason J Meudt, Jessica Strang, Dustin Hebel, Alyx Alfson, Stephanie J Olson, Tera R Kruser, Jennifer B Smilowitz, et al. 2024. Dataset for gastrointestinal tract segmentation on serial mris for abdominal tumor radiotherapy. *Data in Brief*, page 111159.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Yuanpeng Li, Liang Zhao, Jianyu Wang, and Joel Hestness. 2019. Compositional generalization for primitive substitutions. *arXiv preprint arXiv:1910.02612*.

Yuexiang Li, Nanjun He, and Yawen Huang. 2022. Single domain generalization via spontaneous amplitude spectrum diversification. In *MICCAI Workshop on Resource-Efficient Medical Image Analysis*, pages 32–41. Springer.

Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. 2021. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052.

Xiao Liang. 2021. Adam dataset. https://www.kaggle.com/datasets/xiaoliang2121/adamdataset. Kaggle.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Jacob A Macdonald, Zhe Zhu, Brandon Konkel, and Mazurowski. 2020. Siim-acr pneumothorax segmentation. https://doi.org/10.5281/zenodo.7774566. Zenodo.

K Scott Mader. 2017. Mias mammography. https://www.kaggle.com/datasets/kmader/mias-mammography. Kaggle.

Salman Maqbool, Aqsa Riaz, Hasan Sajid, and Osman Hasan. 2020. m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks. *arXiv preprint arXiv:2008.10134*.

Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. 2021. An expert-annotated dataset of bone marrow cytology in hematologic malignancies. *The Cancer Imaging Archive*.

Sarah Matta, Mathieu Lamard, Philippe Zhang, Alexandre Le Guilcher, Laurent Borderie, Béatrice Cochener, and Gwenolé Quellec. 2024. A systematic review of generalization research in medical image classification. *arXiv preprint arXiv:2403.12167*.

Teresa Mendonca, M Celebi, T Mendonca, and J Marques. 2015. Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy image analysis*.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge\-mobile-devices/.

Shentong Mo and Paul Pu Liang. 2024. Multimed: Massively multimodal and multitask medical understanding. *arXiv preprint arXiv:2408.12682*.

Paul Mooney. 2017. Blood cell images. https://www.kaggle.com/datasets/paultimothymooney/blood-cells. Kaggle.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Loris Nanni, Michelangelo Paci, Florentino Luciano Caetano dos Santos, Heli Skottman, Kati Juuti-Uusitalo, and Jari Hyttinen. 2016. Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium. *PLoS One*, 11(2):e0149399.

Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. 2023. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277.

Masoud Nickparvar. 2021a. Brain tumor mri dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset. Kaggle.

Msoud Nickparvar. 2021b. Brain tumor mri dataset.

Nikita Orlov, Wayne Chen, David Eckley, Tomasz Macura, Lior Shamir, Elaine Jaffe, and Ilya Goldberg. 2010a. Automatic classification of lymphoma images with transform-based global features. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 14:1003–13.

Nikita Orlov, Wayne Chen, David Eckley, Tomasz Macura, Lior Shamir, Elaine Jaffe, and Ilya Goldberg. 2010b. Automatic classification of lymphoma images with transform-based global features. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 14:1003–13.

Silvia Ovreiu, Elena-Anca Paraschiv, and Elena Ovreiu. 2021. Deep learning & digital fundus images: Glaucoma detection using densenet. In *2021 13th international conference on electronics, computers and artificial intelligence (ECAI)*, pages 1–4. IEEE.

Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. 2020. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221.

Sachin Panchal, Ankita Naik, Manesh Kokare, Samiksha Pachade, Rushikesh Naigaonkar, Prerana Phadnis, and Archana Bhange. 2023. Retinal fundus multi-disease image dataset (rfmid) 2.0: a dataset of frequently and rarely identified diseases. *Data*, 8(2):29.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

H Hieu Pham, T Thanh Tran, and Ha Quy Nguyen. 2022. Vindr-pcxr: An open, large-scale pediatric chest x-ray dataset for interpretation of common thoracic diseases. *PhysioNet (version 1.0. 0)*, 10:2.

Hieu Huy Pham, H Nguyen Trung, and Ha Quy Nguyen. 2021. Vindr-spinexr: A large annotated medical image dataset for spinal lesions detection and classification from radiographs. *PhysioNet*.

Konstantin Pogorelov, Kristin Ranheim Randel, Thomas de Lange, Sigrun Losada Eskeland, Carsten Griwodz, Dag Johansen, Concetto Spampinato, Mario Taschwer, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017a. Nerthus: A bowel preparation quality video dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 170–174, New York, NY, USA. ACM.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017b. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA. ACM.

Praveen. 2019. Coronahack chest x-ray dataset. https://www.kaggle.com/datasets/praveengovi/coronahack-chest-xraydataset. Kaggle.

Pavle Prentasic, Sven Loncaric, Zoran Vatavuk, Goran Bencic, Marko Subasic, Tomislav Petković, Lana Dujmovic, Maja Malenica Ravlic, Nikolina Budimlija, and Rašeljka Tadić. 2013. Diabetic retinopathy image database(dridb): A new database for diabetic retinopathy screening programs research. In *International Symposium on Image and Signal Processing and Analysis, ISPA*, pages 711–716.

Xianbiao Qi, Guoying Zhao, Jie Chen, and Matti Pietikäinen. 2016. Hep-2 cell classification: The role of gaussian scale space theory as a pre-processing approach. *Pattern Recognition Letters*, 82:36–43.

Raddar. 2019. Chest x-rays (indiana university). https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university?select=indiana_reports.csv. Kaggle.

Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. 2020. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *Ieee Access*, 8:191586–191601.

MOHD ZAID RASHID. 2024. Oral cancer dataset. https://www.kaggle.com/datasets/zaidpy/oral-cancer-dataset. Kaggle.

Sucheng Ren, Xiaoke Huang, Xianhang Li, Junfei Xiao, Jieru Mei, Zeyu Wang, Alan Yuille, and Yuyin Zhou. 2024. Medical vision generalist: Unifying medical imaging tasks in context. *arXiv preprint arXiv:2406.05565*.

Manuel Alejandro Rodríguez, Hasan AlMarzouqi, and Panos Liatsis. 2022. Multi-label retinal disease classification using transformers. *IEEE Journal of Biomedical and Health Informatics*.

Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Salman Sajid. 2024. Oral diseases. https://www.kaggle.com/datasets/salmansajid05/oral-diseases/data. Kaggle.

F Shaker. 2018. Human sperm head morphology dataset (hushem). *Mendeley Data*, 3.

Julio Silva-Rodríguez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. 2020. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine*, 195:105637.

Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. 2020. Sars-cov-2 ct-scan dataset:a large dataset of real patients ct scans for sars-cov-2 identification. *Cold Spring Harbor Laboratory Press*.

Malliga Subramanian, Kogilavani Shanmugavadivel, Obuli Sai Naren, K Premkumar, and K Rankish. 2022. Classification of retinal oct images using deep learning. In *2022 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–7.

Summers and Ronald. 2020. Chestxray nihcc. https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345. NIH.

SunneYi. 2021. Chest CT-Scan images Dataset.

Siham Tabik, Anabel Gómez-Ríos, José Luis Martín-Rodríguez, Iván Sevillano-García, Manuel Rey-Area, David Charte, Emilio Guirado, Juan-Luis Suárez, Julián Luengo, MA Valero-González, et al. 2020. Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images. *IEEE journal of biomedical and health informatics*, 24(12):3595–3605.

Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. 2024. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning. *arXiv preprint arXiv:2410.16162*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

Peking University. 2019. Odir-2019 dataset. https://odir2019.grand-challenge.org/introduction/. Grand Challenge.

Preet Viradiya. 2020. Brain tumor dataset. https://www.kaggle.com/datasets/preetviradiya/brain-tumor-dataset. Kaggle.

Haiyang Wang, Hao Tang, Li Jiang, Shaoshuai Shi, Muhammad Ferjad Naeem, Hongsheng Li, Bernt Schiele, and Liwei Wang. 2025. Git: Towards generalist vision transformer through universal language interface. In *European Conference on Computer Vision*, pages 55–73. Springer.

Linda Wang, Zhong Qiu Lin, and Alexander Wong. 2020. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024c. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

wjXiaochuangw. 2019. Covid-19-ct scan images.

Zhenlin Xu, Marc Niethammer, and Colin A Raffel. 2022. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. *Advances in Neural Information Processing Systems*, 35:25074–25087.

Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. 2019. Siim-acr pneumothorax segmentation. https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation. Kaggle.

Yaya Zha. 2021. Rus-chn. https://aistudio.baidu.com/datasetdetail/69582/0. AI Studio.

Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. 2023a. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*.

Edward Zhang and Sauman Das. 2022. Glaucoma detection. https://www.kaggle.com/datasets/sshikamaru/glaucoma-detection. Kaggle.

Ruipeng Zhang, Qinwei Xu, Chaoqin Huang, Ya Zhang, and Yanfeng Wang. 2022. Semi-supervised domain generalization for medical image analysis. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2023b. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Change Loy Chen, and Shuicheng Yan. 2024a. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In *NeurIPS*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023c. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024b. Why are visually-grounded language models bad at image classification? *arXiv preprint arXiv:2405.18415*.

Jinyu Zhao, Yichen Zhang, Xuehai He, and Pengtao Xie. 2020. Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*.

Chuang Zhu, Wenkai Chen, Ting Peng, Ying Wang, and Mulan Jin. 2021a. Hard sample aware noise robust learning for histopathology image classification.

IEEE transactions on medical imaging, 41(4):881–894.

Chuang Zhu, Wenkai Chen, Ting Peng, Ying Wang, and Mulan Jin. 2021b. Hard sample aware noise robust learning for histopathology image classification. *IEEE transactions on medical imaging*, 41(4):881–894.

Xile Zhu. 2022. Lc25000. https://www.kaggle.com/datasets/xilezhu/lc25000. Kaggle.

. 2021. Augemnted ocular diseases. https://www.kaggle.com/datasets/nurmukhammed7/augemnted-ocular-diseases. Kaggle.

A More Experiments

A.1 Benefits for Segmentation

Segmentation-enabled LLMs, such as Next-GPT, first use the LLM to identify potential regions of the target object and then apply a SAM to decode the object mask, thereby completing the segmentation task. In this context, segmentation can be seen as an extension of detection, potentially requiring more images to achieve improved performance. We conducted additional experiments to explore whether MLLMs can still utilize CG to understand new images across both segmentation and classification tasks.

Related Combination		Target Subset	Next-Chat
D - Skin	C - Intestine	S - Intestine	+7.46
D - Intestine	C - Skin	S - Skin	+5.42

Table 6: *Next-Chat* represents the cIoU gain brought by CG. C indicates classification task, S indicates Segmentation task.

The results in Table 6 demonstrate that, in the context of segmentation tasks, MLLMs are still able to leverage CG to understand new tasks, which is consistent with our original conclusions.

A.2 More Complex Medical Elements

While MAT-Triplet Categorization is useful, predefined categories may limit the exploration of more complex medical attributes, so we also considered integrating more flexible categorization to explore additional medical attributes.

Additional Element 1: Population Groups We selected VinDr-PCXR and MedMAT Subset 31 for the experiment, as they contain X-ray images of children and adult groups, respectively. The results are shown in Table 7.

Additional Element 2: Finer Disease "Finer disease" means more detailed categorization. For instance, we treat COVID and common pneumonia as distinct diseases for generalization. We split the Normal data in the training set into two parts and combined each with COVID and Pneumonia data to create new datasets. The results are shown in Table 8.

Related Combination		Target Subset	LLaVA
X-ray Young	Unrelated Data	X-ray Adults	+6.04
X-ray Young	CT Children (CG)	X-ray Adults	+18.12

Table 7: Results of using *Population Groups* as a CG element.

Related Combination		Target Subset	LLaVA
X-ray Pneumonia	Unrelated Data	X-ray COVID	+11.33
X-ray Pneumonia	CT COVID (CG)	X-ray COVID	+12.67

Table 8: Results of using *Finer Disease* as a CG element.

A.3 Statistical Tests of the Generalization Results

To ensure consistency and repeatability of the experiment, we performed statistical tests in this section. LLaVA is selected as the baseline, and we used the same data combinations from Section 3.3. Each experiment was repeated 3 times, and we reported the mean and standard deviation (SD) of the results.

From the results in Table 9, we can observe that the outcomes across runs show low variance, indicating overall stability, and they continue to support our original experimental conclusions.

A.4 CG with All MAT-Triplet Elements from Different Sources

In previous controlled experiments (Section 3), one element of the MAT-Triplet was kept constant while CG was explored in the remaining two elements. To ensure that all the 3 MAT-Triplet elements of the target data originated from three distinct datasets, additional experiments were conducted to further validate the effectiveness of CG. For these experiments, all possible combinations meeting the criteria in Med-MAT were selected (**Selection Strategy**). The results presented in Table 10 demonstrate that most combinations can effectively generalize to the *Target* data.

Analysis of the results The results in Table 7 and 8 indicate that the two new attributes show data leakage due to subtle visual differences in corresponding images (e.g., COVID-19 and pneumonia

Related Combination				Target Subset		Baseline	1st	2nd	3rd	Mean and SD
Bones	State	Breast	**Diseases**	**Bones**	**Diseases**	37.31	43.28	44.78	43.28	43.78 ± 0.87
Lung	COVID	Bones	**Diseases**	**Lung**	**Diseases**	49.00	52.00	52.00	52.00	52.00 ± 0.00
X-ray	Diseases	CT	**COVID**	**X-ray**	**COVID**	30.00	47.33	49.33	49.33	48.66 ± 1.15
X-ray	Diseases	CT	**State**	**X-ray**	**State**	30.00	46.00	45.33	44.67	45.33 ± 0.67
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	25.00	31.50	32.00	32.00	31.83 ± 0.29
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	47.00	71.00	71.00	70.00	70.67 ± 0.58
FP - Fundus	Diseases	Mic - Cell	**Level**	**FP - Fundus**	**Level**	33.33	42.42	45.45	45.45	44.44 ± 1.75
Mic - Cell	Recognition	FP - Fundus	**Level**	**Mic - Cell**	**Level**	23.00	32.00	32.00	31.50	31.83 ± 0.29

Table 9: Statistical tests of CG experiments. The 1st, 2nd, and 3rd show the generalization results of the experiment in different runs. "Mean" and "SD" represent the average accuracy (%) and standard deviation.

Related Combination			Target Subset			Baseline	Trained	CG Helps
CT	Brain	Cancer	CT	Brain	Cancer	28	26	✗
CT	Brain	Cancer	CT	Brain	Cancer	28	25	✗
CT	Brain	State	CT	Brain	State	33	64	✓
CT	Brain	State	CT	Brain	State	33	70	✓
X-ray	Lung	Diseases	X-ray	Lung	Diseases	30	45	✓
X-ray	Lung	Diseases	X-ray	Lung	Diseases	30	38	✓
X-ray	Lung	Diseases	X-ray	Lung	Diseases	30	44	✓
X-ray	Breast	Diseases	X-ray	Breast	Diseases	31	32	✓
X-ray	Breast	Diseases	X-ray	Breast	Diseases	31	52	✓

Table 10: Results from 3 datasets providing different elements of MAT-Triplet. ✓ in *CG Helps* indicates successful generalization, while ✗ denotes failure.

have similar features). Importantly, the MLLM trained with CG combinations still shows improvements on downstream tasks, confirming that our approach remains valid for new attributes.

Reason to choose the existing three attributes (MAT-Triplet: Modality, Area, Task) We have considered additional categories such as age, gender, and finer disease classification, but we ultimately chose to focus on the MAT-Triplet categories for the following reasons.

- **The boundaries between MAT-Triplet (Modality, Area, Task) are clear.** Different modalities and areas correspond to distinct imaging methods and body areas, leading to significant differences between images; different tasks also require the MLLM to extract specific information, demanding varied understanding of the images.
- **All datasets can be annotated using MAT-Triplet (Modality, Area, Task) easily.** Other medical labels, such as gender and age, are only available in a small portion of datasets and are not suitable for large-scale annotation.
- **Similar categorization strategies have been adopted in previous studies.**

A.5 Details of Section 3.3: Exploring CG on different MLLM Backbones

To ensure the experiment results are not influenced by the model choice, we also tested several other models on some subsets of Med-MAT and observed similar results.

Selection Strategy: For testing, some generalized combinations were selected from classification tasks 3. Using a random seed of 42, we shuffled each Direction Type's combinations and selected the first two compositions as test data.

Experimental Setup: We conducted experiments to evaluate the compatibility of CG across different backbone architectures. We selected two MLLMs with representative architectures, namely Qwen2-VL-7B-Instruct (Wang et al., 2024b) and Llama-3.2-11B-Vision-Instruct (Meta AI, 2024), to assess the performance of CG on these models. Each experiment involved full-parameter fine-tuning of all models over 5 epochs, utilizing 8 A800 (80GB) GPUs. The training was performed with a batch size of 32 and a learning rate set to 2e-6, ensuring that all parameters were updated to optimize the model performance.

A.6 Details of Section 6: Exploring CG across Detection and Classification

Experimental Setup: We conducted generalization experiments for detection and classification. Specifically, we performed generalization validation on Next-Chat (Zhang et al., 2023a) and MiniGPT-v2 (Chen et al., 2023a). Next-Chat models the bounding box as an embedding and utilizes a decoder for decoding, while MiniGPT-v2 treats the bounding box as a text token, which are common approaches used by existing MLLM implementations for detection. By conducting CG validation using distinct bounding box modeling methods, we further demonstrate the broad applicability of the CG approach. Each experiment was conducted on 8 A800 (80GB) GPUs.

The two backbones were trained separately in this experiment. For Next-Chat, we directly trained the model in its second training stage and fine-tuned it for 2 epochs with a learning rate of 2e-5,

Related Combination				Target Subset		Baseline	Trained	CG Helps
Bones	State	Breast	**Diseases**	**Bones**	**Diseases**	61	65	✓
Lung	COVID	Bones	**Diseases**	**Lung**	**Diseases**	80	91	✓
X-ray	Diseases	CT	**COVID**	**X-ray**	**COVID**	35	40	✓
X-ray	Diseases	CT	**State**	**X-ray**	**State**	35	43	✓
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	32	33	✓
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	65	72	✓
FP - Fundus	Diseases	Mic - Cell	**Level**	**FP - Fundus**	**Level**	48	45	✗
Mic - Cell	Recognition	FP - Fundus	**Level**	**Mic - Cell**	**Level**	34	41	✓

Table 11: Result of Qwen2-VL on selected classification datasets in Med-MAT. ✓ in *CG Helps* indicates successful generalization, while ✗ denotes failure.

Related Combination				Target Subset		Baseline	Trained	CG Helps
Bones	State	Breast	**Diseases**	**Bones**	**Diseases**	52	59	✓
Lung	COVID	Bones	**Diseases**	**Lung**	**Diseases**	64	75	✓
X-ray	Diseases	CT	**COVID**	**X-ray**	**COVID**	33	38	✓
X-ray	Diseases	CT	**State**	**X-ray**	**State**	33	41	✓
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	31	29	✗
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	49	57	✓
FP - Fundus	Diseases	Mic - Cell	**Level**	**FP - Fundus**	**Level**	55	61	✓
Mic - Cell	Recognition	FP - Fundus	**Level**	**Mic - Cell**	**Level**	10	32	✓

Table 12: Result of Llama-3.2-Vision on selected classification datasets in Med-MAT. ✓ in *CG Helps* indicates successful generalization, while ✗ denotes failure.

Related Combination				Target Subset		Baseline	Trained	CG Helps
Lung	Lung Det	Bones	**Diseases**	**Lung**	**Diseases**	49	52	✓
Lung	Lung Det	Breast	**Diseases**	**Lung**	**Diseases**	49	54	✓
Bones	Spinal Error Det	Breast	**Diseases**	**Bones**	**Diseases**	20	30	✓
Bones	Spinal Error Det	Lung	**Diseases**	**Bones**	**Diseases**	20	33	✓
End	Level	MRI	**Diseases Det**	**End**	**Diseases**	24	27	✓
X-ray	Lung Det	CT	**COVID**	**X-ray**	**COVID**	23	26	✓
Der - Skin	Cancer Det	FP - Fundus	**Diseases**	**Der - Skin**	**Diseases**	24	29	✓
Mic - Cell	Cancer Det	CT - Kidney	**Diseases**	**Mic - Cell**	**Diseases**	24	26	✓

Table 13: Result of NEXT-Chat on CG by using detection and classification tasks to generalize classification Target dataset. Generalization results on classification datasets: *Related Combination* is the training set, *Target Subset* is the goal. Baseline and Trained represent the model's accuracy without training and trained on related data, respectively. ✓ in *CG Helps* indicates successful generalization, while ✗ denotes failure.

Related Combination				Target Subset		Baseline	Trained	CG Helps
Lung	Lung Det	Bones	**Diseases**	**Lung**	**Diseases**	41	47	✓
Lung	Lung Det	Breast	**Diseases**	**Lung**	**Diseases**	41	49	✓
Bones	Spinal Error Det	Breast	**Diseases**	**Bones**	**Diseases**	31	35	✓
Bones	Spinal Error Det	Lung	**Diseases**	**Bones**	**Diseases**	31	37	✓
End	Level	MRI	**Diseases Det**	**End**	**Diseases**	24	26	✓
X-ray	Lung Det	CT	**COVID**	**X-ray**	**COVID**	22	23	✓
Der - Skin	Cancer Det	FP - Fundus	**Diseases**	**Der - Skin**	**Diseases**	27	30	✓
Mic - Cell	Cancer Det	CT - Kidney	**Diseases**	**Mic - Cell**	**Diseases**	20	24	✓

Table 14: Result of MiniGPT-v2 on CG by using detection and classification tasks to generalize classification Target dataset. Generalization results on classification datasets: *Related Combination* is the training set, *Target Subset* is the goal. Baseline and Trained represent the model's accuracy without training and trained on related data, respectively. ✓ in *CG Helps* indicates successful generalization, while ✗ denotes failure.

keeping all other training parameters at their default settings. Similarly, for MiniGPT-v2, we trained the backbone model from the second stage, starting with a learning rate of 2e-5 and gradually reducing it to 2e-6 over 3 epochs.

A.7 CG with Medical Multimodal LLM

In previous experiments, general MLLMs are selected to prevent the MLLM's inherent medical knowledge from affecting CG results. Our experiments focus on how MLLMs leverage CG to interpret unseen medical images. If the model has

Related Combination				Target Subset		HuatuoGPT
Bones	State	Breast	**Diseases**	**Bones**	**Diseases**	+6.12
Lung	COVID	Bones	**Diseases**	**Lung**	**Diseases**	+15.00
X-ray	Diseases	CT	**COVID**	**X-ray**	**COVID**	+38.00
X-ray	Diseases	CT	**State**	**X-ray**	**State**	+40.67
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	+1.5
CT	Brain	X-ray	**Lung**	**CT**	**Lung**	+18.00
FP - Fundus	Diseases	Mic - Cell	**Level**	**FP - Fundus**	**Level**	+12.12
Mic - Cell	Recognition	FP - Fundus	**Level**	**Mic - Cell**	**Level**	+10.50

Table 15: Result of HuatuoGPT-Vision on selected classification datasets in Med-MAT. *HuatuoGPT* represent the accuracy(%) gains the model achieved through CG.

learned some fundamental elements of the *Target* data, it would compromise the fairness of the experiments.

To demonstrate that our results still work on medical LLMs, we employed the same data combinations from Section 3.3 to investigate CG on medical MLLMs (we selected HuatuoGPT-Vision as the baseline).

The results in Table 15 demonstrate that **the medical-expert MLLM can still leverage CG to enhance their performance on novel tasks**, further supporting the validity and consistency of our findings.

B The Dataset: Med-MAT

This section provides an overview of Med-MAT. First, a detailed explanation of MAT-Triplet will be presented in B.1. Next, the methods for constructing the QA formatting will be discussed in B.2. Finally, the data composition details and open-source specification will be provided in B.3.

B.1 Details of MAT-Triplet

MAT-Triplet stands for **M**edical Modality, **A**natomical Area, and Medical **T**ask. We define all samples in Med-MAT using these three components and integrate datasets with identical triplets into subsets.

Medical Modality refers to different types of techniques or methods used in medical imaging or data acquisition. Each modality is designed to present the human body's structures or pathological features in unique ways, providing auxiliary support for clinical diagnosis and treatment. Most modalities exhibit significant visual differences, making them easily distinguishable. Med-MAT encompasses 11 modalities, including common ones such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), X-ray, Fundus Photography (FP), Endoscopy (End), Optical Coherence Tomography (OCT), and Ultrasound (US), as well as rare and specialized modalities like Colonoscopy (Co), Dermoscopy (Der), Digital Pathology (DP), and Microscopy (Mic).

Anatomical Area refers to specific anatomical structures or regions within the human body or other organisms, defined by distinct anatomical characteristics to describe various body parts, their functions, and relative positions. Med-MAT encompasses 14 anatomical areas, including the cervix, kidney, lung, brain, intestine, bladder, fundus, retina, breast, bones, and chest. To further facilitate data description, additional categories such as skin, mouth, and cell are included as specialized anatomical areas.

Medical Task refers to the specific detection task that needs to be performed on the dataset. Med-MAT includes 13 distinct tasks, with classification tasks encompassing Quality Identification (image quality analysis), COVID Diagnosis, Cancer Diagnosis (determining the presence of a specific disease), State (such as identifying brain hemorrhage), Level Identification (assessing disease severity), and Multiple Classification (classifying multiple diseases or cell types). Given the limited options of COVID Diagnosis and Cancer Diagnosis, these tasks can be interpreted as identifying whether a patient is in a diseased state. To enhance generalization and provide more diverse examples, these tasks are grouped under the broader category of State. In addition, we have 16 datasets defining segmentation or classification tasks with different objectives.

B.2 QA construction method

A large amount of image-label datasets was collected to build the Med-MAT dataset. To ensure compatibility with MLLM training inputs and outputs, all data is transformed into a question-answering format. Questions are formulated based

on modality, anatomical area, and medical task, with 6 question prompts applied to each subset.

The labels within each data subset will be clustered to prevent redundant definitions of the same condition. Then, all training set and test set will be converted into multiple-choice questions following the template in Table 8. Each question will have up to four options, with distractor options randomly selected from the corresponding subset.

B.3 Data composition and Open-source Specification

Med-MAT is composed of multiple datasets. After being transformed into different QA formats, the new data is organized into several subsets to support generalization experiments in medical imaging. Table 17 shows all of our subset datasets, which are separated based on different combinations in MAT-Triplet. The specific MAT-Triplets are listed, along with the labels corresponding to the image-label datasets for each subset. Correspondingly, all the image-label datasets are also displayed in Table 18, which includes their names, descriptions of the tasks performed, download links, and the level of accessibility.

All question-answering text datasets in Med-MAT will be publicly available. To accommodate varying access permissions, we will release datasets based on their respective licenses: openly accessible datasets will be directly available, while restricted datasets can be accessed by applying through the links provided in this paper. We hope this dataset will support and advance future generalization experiments on medical imaging.

B.4 Data Sources and Distribution

All Med-MAT data are sourced from public medical image challenges or widely used, high-impact datasets previously applied in deep learning training, ensuring reliable annotations. Before inclusion in Med-MAT, all datasets underwent label averaging where possible; test sets, in particular, were strictly balanced to ensure accuracy reliably reflects model performance. Each Med-MAT training subset contains 3,000 samples, while test sets maximize size under label balance constraints.

C Bad cases analysis and solutions

C.1 Bad case analysis

Some *Trained* models show minimal gains or even performance declines in Table 3, with classification accuracy lower than either the *Baseline* or *Baseline+*. After a thorough examination, we found that these Target datasets require more fine-grained medical condition classification. Beyond disease presence, they need detailed assessments, such as severity grading (e.g., bone age estimation, cancer staging) or distinguishing similar conditions (e.g., differentiating COVID-19 from pneumonia).

- The *Related* combinations lack suitable fundamental elements: For CG, the training data must include the *Target* task's core elements. Here, we use other "level classification/grading" tasks for generalization, but their criteria differ significantly, misaligning with the *Target* task's needs.
- Without defined grading standards, MLLMs lacking relevant knowledge can't perform fine-grained tasks: Tasks like bone age assessment and cancer staging vary by criteria, and without this knowledge, MLLMs can't accurately classify them.

C.2 Possible solutions

Few-shot prompting As we illustrated before, most of the bad cases involve fine-grained tasks needing specialized knowledge. So, in order to minimize the effect of a lack of relevant knowledge, we also conducted few-shot experiments to add some target images in the prompts. Subset *X-ray, Lung, Normal-COVID-Pneumonia* was chosen for its simple structure, with LLaVA as the baseline. We randomly sampled n images per label for n-shot inference and repeated each experiment 3 times.

Model	0-shot	2-shot	3-shot	4-shot
LLaVA	30.00	**28.83** $\pm$ 0.85	29.33 $\pm$ 1.25	29.83 $\pm$ 1.31
LLaVA + CG	28.00	28.67 $\pm$ 0.94	**37.00** $\pm$ 0.82	**36.67** $\pm$ 0.47

Table 16: Results of Few-shot prompting.

The results in Table 16 demonstrate that training with CG combinations can improve the few-shot performance of MLLMs on downstream tasks, even when direct CG generalization is not effective.

Adding some *Target* data in training As described in Section 5.2, we selected cases where CG alone couldn't achieve satisfactory results and augmented their training sets with target data. The results in this section indicate that while CG may not directly enhance generalization, it accelerates the model's adaptation to downstream tasks.

Multiple-choice Questions Template

<question>
A. <option_1>
B. <option_2>
C. <option_3>
D. <option_4>
Answer with the option's letter from the given choices directly.

Figure 8: The Template of multiple-choice questions.

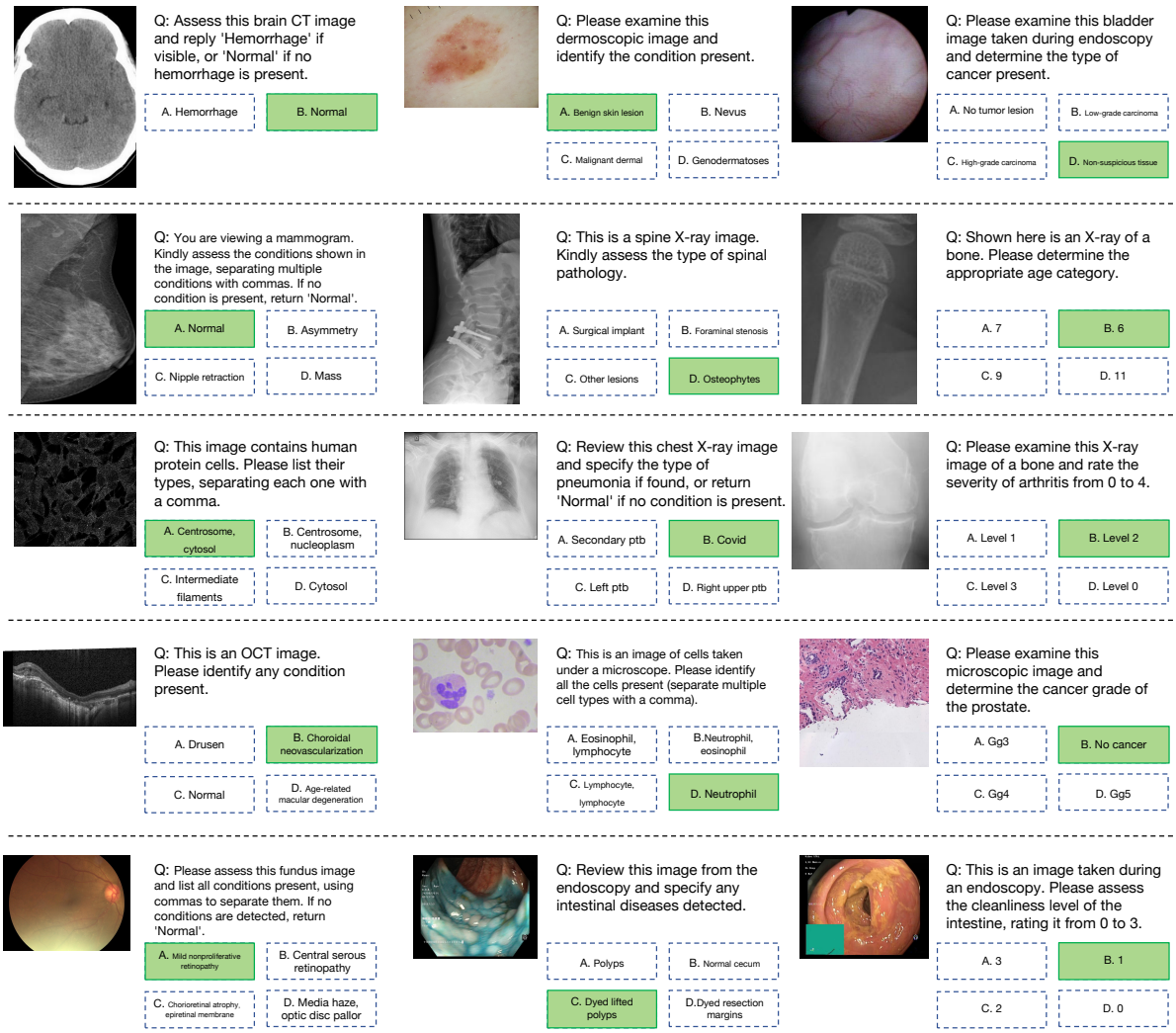

Figure 9: Illustration of diverse samples with varying numbers of candidate options in the Med-MAT dataset.

Subset No.	Modality	Anatomical Area	Task	Datasets No.
01	Co	Cervix	Cervical Picture Quality Evaluation	1
02	CT	Kidney	Kidney Diseases Classification	2
03	CT	Lung	COVID-19 Classification	3,4,6
04	CT	Lung	Lung Cancer Classification	5
05	CT	Brain	Brain Hemorrhage Classification	7
06	CT	Brain	Brain Cancer Classification	8
07	Der	Skin	Melanoma Type Classification	10
08	Der	Skin	Skin Diseases Classification	9, 11-15, 71, 72, 74
09	DP	Mouth	Teeth Condition Classification	16
10	DP	Mouth	Oral Cancer Classification	17
11	End	Intestine	Intestine Cleanliness Level	18
12	End	Bladder	Cancer Degree Classification	19
13	End	Intestine	Intestine Diseases Classification	20
14	FP	Fundus	Eye Diseases Classification	21-23, 26-28, 31, 32, 75
15	FP	Fundus	Multiple-labels Eye Diseases Classification	24, 25, 68
16	FP	Fundus	Blindness Level	29
17	FP	Fundus	Retinal Images Quality Evaluation	30
18	Mic	Cell	Cell Type Classification	33, 36-38, 39-41, 44, 65, 70
19	Mic	Cell	Prostate Cancer Degree Classification	34
20	Mic	Cell	Multiple-labels Blood Cell Classification	35
21	Mic	Cell	Cancer Classification	42, 67
22	MRI	Brain	Head Diseases Classification	44, 45
23	OCT	Retina	Retina Diseases Classification	46, 47
24	US	Breast	Breast Cancer Classification	48
25	X-ray	Bones	Degree Classification of Knee	49, 53
26	X-ray	Bones	Fractured Classification	50, 51
27	X-ray	Bones	Vertebrae Diseases Classification	52
28	X-ray	Lung	COVID-19 and Pneumonia Classification	54-57, 60, 62, 81
29	X-ray	Breast	Breast Diseases Classification	58, 78
30	X-ray	Lung	Tuberculosis Classification	59, 79
31	X-ray	Chest	Multiple-labels Chest Classification	61, 73, 76, 77, 80, 85, 87
32	X-ray	Brain	Tumor Classification	63
33	Mic	Cell	Multi-labels Diseases	84
34	FP	Fundus	Level Identification	66
35	X-ray	Bones	Level Identification	69
36	X-ray	Bones	Spinal lesion Classification	86
37	X-ray	Breast	Multi-labels Diseases	82
38	Der	Skin	Lesion Det/Seg	88-91
39	End	Intestine	PolyP Det/Seg	92-93
40	End	Intestine	Surgical Procedures Det/Seg	94
41	End	Intestine	Multi-labels Det/Seg	95
42	Mic	Cell	Cancer Cell Det/Seg	96
43	US	Chest	Cancer Det/Seg	97
44	US	Thyroid	Thyroid Nodule Region Det/Seg	98
45	MRI	Intestine	Multi-labels Det/Seg	103
46	MRI	Liver	Liver Det/Seg	104, 105
47	X-ray	Lung	Lung Det/Seg	99
48	X-ray	Lung	Pneumothorax Det/Seg	106
49	X-ray	Bones	Spinal Anomaly Det	100
50	X-ray	Chest	Multi-labels Det	101, 102
51	FP	Fundus	Vessel Seg	107
52	FP	Fundus	Optic Disc and Cup Seg	108
53	FP	Fundus	Optic Disc Seg	109

Table 17: The details of subset. In particular, **Co** stands for Colposcopy, **CT** represents Computed Tomography, **DP** refers to Digital Photography, **FP** is for Fundus Photography, **MRI** denotes Magnetic Resonance Imaging, **OCT** signifies Optical Coherence Tomography, **Der** refers to Dermoscopy, **End** stands for Endoscopy, **Mic** indicates Microscopy Images, and **US** represents Ultrasound. The blue section represents the classification dataset and the green section represents the detection

No.	Name	Description	Citation
1	Intel & MobileODT Cervical Screening	Cervix Type in Screening	(BenO et al., 2017)
2	CT Kindney Dataset	Normal or Cyst or Tumor	(Islam et al., 2022a)
3	SARS-COV-2 Ct-Scan	COVID19, Classification Dataset	(Soares et al., 2020)
4	COVID CT COVID-CT	COVID19, Classification Dataset	(Zhao et al., 2020)
5	Chest CT-Scan	Cancer Classification	(SunneYi, 2021)
6	COVID-19-CT SCAN IMAGES	COVID19, Classification	(wjXiaochuangw, 2019)
7	Head CT	Head Hemorrhage	(Kitamura, 2018)
8	CT of Brain	Head Cancer	(Data, 2023)
9	MED-NODE	Melanoma or Naevus	(Giotis et al., 2015)
10	ISIC 2020	Melanoma, Benign or Malignant	(Rotemberg et al., 2021)
11	PAD-UFES-20	Skin Multi Classification	(Pacheco et al., 2020)
12	Web-scraped Skin Image	Skin Desease Multi Classification	(Islam et al., 2022b)
13	ISBI 2016	Skin Lesion Classification	(Gutman et al., 2016)
14	ISIC 2019	Skin Desease Multi Classification	(Combalia et al., 2019)
15	Skin Cancer ISIC	Skin Cancer Multi Classification	(Katanskiy, 2019)
16	Dental Condition Dataset	Teeth condition classification	(Sajid, 2024)
17	Oral Cancer Dataset	Oral cancer Classification	(RASHID, 2024)
18	The Nerthus Dataset	Cleanliness level	(Pogorelov et al., 2017a)
19	Endoscopic Bladder Tissue	Canser Degree Classification	(Lazo et al., 2023)
20	Kvasir	Multi Disease Classification	(Pogorelov et al., 2017b)
21	ACRIMA	Glaucoma	(Ovreiu et al., 2021)
22	Augemnted ocular diseases AOD	Multi Classification of eye diseases	(, 2021)
23	JSIEC	Multi Classification of eye diseases	(Cen et al., 2021)
24	Multi-Label Retinal Diseases	Multi Classification of eye diseases	(Rodríguez et al., 2022)
25	RFMiD 2.0	Multi Classification of eye diseases	(Panchal et al., 2023)
26	ToxoFundus(Data Processed Paper)	Ocular toxoplasmosis	(Cardozo et al., 2023)
27	ToxoFundus(Data Raw 6class All)	Ocular toxoplasmosis	(Cardozo et al., 2023)
28	Adam dataset	Age-related Macular Degeneration	(Liang, 2021)
29	APTOS 2019 Blindness	Blindness Level Identification	(Karthik et al., 2019)
30	DRIMDB	Quality Testing of Retinal Images	(Prentasic et al., 2013)
31	Glaucoma Detection	Glaucoma Classification	(Zhang and Das, 2022)
32	AIROGS	Glaucoma Classification	(de Vente et al., 2023)
33	ICPR-HEp-2	Multi Classification	(Qi et al., 2016)
34	SICAPv2	Cancer Degree Classification	(Silva-Rodríguez et al., 2020)
35	Blood Cell Images	Blood Cell Classificaion	(Mooney, 2017)
36	BreakHis	Cell type and beginormag	(Bukun, 2019)
37	Chaoyang	Multi Classification of pathologists	(Zhu et al., 2021a)
38	HuSHeM	Sperm Head Morphology Classificaion	(Shaker, 2018)
39	Bone Marrow Cell Classification	Bone Marrow Cell Classification	(Matek et al., 2021)
40	NCT-CRC-HE-100K	Multi Classification	(Kather et al., 2018)
41	Malignant Lymphoma Classification	Multi Classification	(Orlov et al., 2010a)
42	Histopathologic Cancer Detection	Cancer Classification	(Cukierski, 2018)
43	LC25000	Multi Classification of Lung and Colon	(Zhu, 2022)
44	Brain Tumor 17 Classes	Multi Classification	(Feltrin, 2022)
45	Tumor Classification	Pituitary or Glioma or Meningioma or Notumor	(Nickparvar, 2021a)
46	Malignant Lymphoma Classification	Multi Classification of eye diseases	(Orlov et al., 2010b)
47	Retinal OCT-C8	Multi Classification of eye diseases	(Subramanian et al., 2022)
48	BUSI	Breast Cancer	(Al-Dhabyani et al., 2020)
49	Digital Knee X-Ray Images	Degree Classification of Knee	(Gornale and Patravali, 2020)
50	Bone Fracture Multi-Region X-ray Data	Fractured Classification	(Nickparvar, 2021b)
51	Fracture detection	Fractured Classification	(Batra, 2024)
52	The vertebrae X-ray image	Vertebrae	(Fraiwan et al., 2022)
53	Knee Osteoarthritis Dataset	Knee Osteoarthritis with severity grading	(Chen, 2018)
54	Shenzhen Chest X-Ray Set	COVID19, Classification Dataset	(Jaeger et al., 2014)
55	Chest X-ray PD	COVID and Pneumonia	(Asraf and Islam, 2021)
56	COVID-19 CHEST X-RAY DATABASE	COVID and Pneumonia	(Chowdhury et al., 2020)
57	COVIDGR	COVID19, Classification	(Tabik et al., 2020)
58	MIAS	Multi Classification of Breast	(Mader, 2017)
59	Tuberculosis Chest X-Ray Database	Tuberculosis	(Rahman et al., 2020)
60	Pediatric Pneumonia Chest X-Ray	Pneumonia Classification	(Kermany, 2018)

Table 18: The details of the medical datasets are provided

No.	Name	Description	Citation
61	Random Sample of NIH Chest X-Ray Dataset	Multi Classificaiton of Chest	(Wang et al., 2017)
62	CoronaHack-Chest X-Ray	Pnemonia Classifcition with Virus type	(Praveen, 2019)
63	Brain Tumor Dataset	Tumor Classification	(Viradiya, 2020)
64	Fitzpatrick 17k (Nine Labels)	Multi Classification	(Groh et al., 2021)
65	BioMediTech	Multi Classification	(Nanni et al., 2016)
66	Diabetic retinopathy	Diabetic Retinopathy Level	(Benítez et al., 2021)
67	Leukemia	Cancer Classification	(Codella et al., 2019)
68	ODIR-5K	Multiple Labels Classification	(University, 2019)
69	Arthrosis	Bone Age Classification	(Zha, 2021)
70	HSA-NRL	Multi Classification of pathologists	(Zhu et al., 2021b)
71	ISIC 2018 (Task 3)	Multi Classification	(Codella et al., 2019)
72	ISIC 2017 (Task 3)	Multi Classification	(Codella et al., 2018)
73	ChestX-Det	Multi Classification	(Lian et al., 2021)
74	Monkeypox Skin Lesion Dataset	Only Monkeypox	(Ali et al., 2022)
75	Cataract Dataset	Multi Classification	(JR2NGB, 2019)
76	ChestX-rays IndianaUniversity	Multi-label Classification	(Raddar, 2019)
77	CheXpert v1.0 small	Multi-label Classification	(Arevalo, 2020)
78	CBIS-DDSM	Multi Classification	(Lee et al., 2017)
79	NLM-TB	Tuberculosis	(Karaca, 2022)
80	ChestXray-NIHCC	Multi-label Classification	(Summers and Ronald, 2020)
81	COVIDx CXR-4	COVID19, Classification	(Wang et al., 2020)
82	VinDr-Mammo	Multi-label Classification	(Nguyen et al., 2023)
83	PBC dataset normal DIB	Multi Classification	(Acevedo et al., 2020)
84	Human Protein Atlas	Multi-label Classification	(Le et al., 2022)
85	RSNA Pneumonia Detection Challenge 2018	Multi-label Classification	(Anouk Stein et al., 2018)
86	VinDr-SpineXR	Multi Classification of Bones Diseases	(Pham et al., 2021)
87	VinDr-PCXR	Multi-label Classification	(Pham et al., 2022)
88	PH2	Melanoma Segmentation	(Mendonca et al., 2015)
89	ISBI 2016 (Task3B)	Melanoma Segmentation	(Gutman et al., 2016)
90	ISIC 2016 (Task 1)	Melanoma Segmentation	(Gutman et al., 2016)
91	ISIC 2017	Melanoma Segmentation	(Codella et al., 2018)
92	CVC-ClinicDB	Polyp Segmentation	(Bernal et al., 2015)
93	Kvasir-SEG	Polyp segmentation	(Jha et al., 2020)
94	m2caiseg	Surgical Instrument Segmentation	(Maqbool et al., 2020)
95	EDD 2020	Multiple Diseases Segmentation in Intestine	(Ali et al., 2020)
96	SICAPv2	Cancer Cells Segmentation	(Silva-Rodríguez et al., 2020)
97	BUSI	Cancer Segmentation	(Hesaraki, 2022)
98	TN3K	Thyroid Nodule Segmentation	(Gong et al., 2022)
99	NLM-TB	Lung Segmentation (With left or right)	(Gong et al., 2021)
100	VinDr-SpineXR	Spinal X-ray Anaomaly Detection	(Pham et al., 2021)
101	VinDr-PCXR	Multiple Diseases Segmentation in Chest	(Pham et al., 2022)
102	ChestX-Det	Multiple Diseases Segmentation in Chest	(Lian et al., 2021)
103	UW-Madison Gl Tract Image Segmentation	Surgical Instrument Segmentation	(Lee et al., 2024)
104	Duke Liver Dataset MRI v1	Liver Segmentation	(Macdonald et al., 2020)
105	Duke Liver Dataset MRI v2	Liver Segmentation	(Macdonald et al., 2020)
106	SIIM-ACR Pneumothorax Segmentation	Pneumothorax Segmentation	(Zawacki et al., 2019)
107	FIVES	Fundus Vascular Segmentation	(Jin et al., 2022)
108	RIM-ONE DL	Optic Disc and Cup Segmentation	(Batista et al., 2020)
109	PALM19	Optic Disc Segmentation	(Fu et al., 2019)

Table 19: Continued from Table 18.